A/B Testing the Udacity Website

IDS 701: Solving Problems with Data

Tea Tafaj Tonantzin Real Rojas

2026-01-27

In these exercises, we'll be analyzing data on user behavior from an experiment run by Udacity, the online education company. More specifically, we'll be looking at a test Udacity ran to improve the onboarding process on their site.

Udacity's test is an example of an "A/B" test, in which some portion of users visiting a website (or using an app) are randomly selected to see a new version of the site. An analyst can then compare the behavior of users who see a new website design to users seeing their normal website to estimate the effect of rolling out the proposed changes to all users. While this kind of experiment has it's own name in industry (A/B testing), to be clear it's just a randomized experiment, and so everything we've learned about potential outcomes and randomized experiments apply here.

(Udacity has generously provides the data from this test under an Apache open-source license, and you can find their original writeup here. If you're interested in learning more on A/B testing in particular, it seems only fair while we use their data to flag they have a full course on the subject here.)

Udacity's Test

The test is described by Udacity as follows:

At the time of this experiment, Udacity courses currently have two options on the course overview page: "start free trial", and "access course materials".

Current Conditions Before Change

- If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first.

- If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

Description of Experimented Change

- In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course.
- If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free.
- At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead. This screenshot shows what the experiment looks like.

Udacity's Hope is that...:

> this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time – without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

Gradescope Autograding

Please follow all standard guidance for submitting this assignment to the Gradescope autograder, including storing your solutions in a dictionary called `results` and ensuring your notebook runs from the start to completion without any errors.

For this assignment, please name your file `exercise_abtesting.ipynb` before uploading.

You can check that you have answers for all questions in your `results` dictionary with this code:

```
assert set(results.keys()) == {
    "ex4_avg_oec",
    "ex5_avg_guardrail",
    "ex7_ttest_pvalue",
    "ex9_ttest_pvalue_clicks",
    "ex10_num_obs",
    "ex11_guard_ate",
```

```
    "ex11_guard_pvalue",
    "ex11_oec_ate",
    "ex11_oec_pvalue",
    "ex14_se_treatment",
}
```

Submission Limits

Please remember that you are **only allowed THREE submissions to the autograder.** Your last submission (if you submit 3 or fewer times), or your third submission (if you submit more than 3 times) will determine your grade Submissions that error out will **not** count against this total.

Import the Data

```
import pandas as pd
import numpy as np
from scipy import stats
import statsmodels.formula.api as smf

# quarto preview exercise_abtesting.ipynb --to pdf
# black exercise_abtesting.ipynb

results = {}
```

Exercise 1

Begin by importing Udacity's data on user behavior here.

There are TWO datasets for this test — one for the control data (users who saw the original design), and one for treatment data (users who saw the experimental design). Udacity decided to show their test site to 1/2 of visitors, so there are roughly the same number of users appearing in each dataset (though this is not a requirement of AB tests).

Please remember to load the data directly from github to assist the autograder.

```
control_path = "https://github.com/nickeubank/MIDS_Data/raw/refs/heads/master/udacity_AB_tes
tmt_path = "https://github.com/nickeubank/MIDS_Data/raw/refs/heads/master/udacity_AB_testing/

df_control = pd.read_csv(control_path)
```

```
df_tmt = pd.read_csv(tmt_path)

print(f"Control: {df_control.shape}")
print(f"Treatment: {df_tmt.shape}")
```

```
Control: (37, 5)
Treatment: (37, 5)
```

Exercise 2

Explore the data. Can you identify the unit of observation of the data (e.g. what is represented by each row)?

To be clear, the columns represent stages in a user funnel:

- Some number of users arrive at the website and are counted as Pageviews,
- Some portion of those users then click to enroll (and are counted as clicks),
- Some portion of those users then actually enroll in the free trial (after seeing an informational popup, in the case of treatment individuals),
- Finally some portion of those users end up paying at the end of the free trial period.

(Note this is not the only way that A/B test data can be collected and/or reported — this is just what Udacity provided, presumably to help address privacy concerns.)

```
print("Control")
df_control.head(3)
```

```
Control
```

	Date	Pageviews	Clicks	Enrollments	Payments
0	Sat, Oct 11	7723	687	134.0	70.0
1	Sun, Oct 12	9102	779	147.0	70.0
2	Mon, Oct 13	10511	909	167.0	95.0

```
print("Treatment")
df_tmt.head(3)
```

```
Treatment
```

	Date	Pageviews	Clicks	Enrollments	Payments
0	Sat, Oct 11	7716	686	105.0	34.0
1	Sun, Oct 12	9288	785	116.0	91.0
2	Mon, Oct 13	10480	884	145.0	79.0

Answer

The unit of observation is the **date**. For each date, we observe the number of users who arrived at the webpage (pageviews), the number who clicked to enroll, the number who enrolled in the free trial, and the number who ultimately paid for the service.

Pick your measures

Exercise 3

The easiest way to analyze this data is to stack it into a single dataset where each observation is a day-treatment-arm (so you should end up with two rows per day, one for those who are in the treated groups, and one for those who were in the control group). Note that currently nothing in the data identifies whether a given observation is a treatment group observation or a control group observation, so you'll want to make sure to add a "treatment" indicator variable.

The variables in the data are:

- Pageviews: number of unique users visiting homepage
- Clicks: number of those users clicking "Start Free Trial"
- Enrollments: Number of people enrolling in trial
- Payments: Number of people who eventually pay for the service. Note the `payment` column reports payments for the users who first visited the site on the reported date, not payments occurring on the reported date.

```
df_control["Treatment"] = 0
df_tmt["Treatment"] = 1

df = pd.concat([df_control, df_tmt])
print("Shape:", df.shape)
df.head(3)
```

```
Shape: (74, 6)
```

	Date	Pageviews	Clicks	Enrollments	Payments	Treatment
0	Sat, Oct 11	7723	687	134.0	70.0	0
1	Sun, Oct 12	9102	779	147.0	70.0	0
2	Mon, Oct 13	10511	909	167.0	95.0	0

```
df.groupby("Treatment")["Date"].count()
```

```
Treatment
0    37
1    37
Name: Date, dtype: int64
```

Exercise 4

Given Udacity's goals, what outcome are they hoping will be impacted by their manipulation?

Or, to ask the same question in the language of the Potential Outcomes Framework, what is their Y?

Or to ask the same question in the language of Kohavi, Tang and Xu, what is their *Overall Evaluation Criterion (OEC)*?

(I'm only asking one question, I'm just trying to phrase it using different terminologies we've encountered to help you see how they all fit together)

When you feel like you have your answer, please compute it. Store the average value of the variable in `results` under the key `ex4_avg_oec`. **Please round your answer to 4 decimal places.**

NOTE: You'll probably notice you have two choices to make when it comes to actually computing the OEC.

- You could probably imagine either computing a ratio or a difference of two things — please calculate the difference.
- You may also be unsure whether to normalize by `Clicks`. Normalizing by clicks will help account for variation that comes from day-to-day variation in users, so it's a good thing to do. With infinite data, you'd expect to get the same results without normalizing by `Clicks` (since on average the same share of users are in each arm of the experiment), but for finite data it's a good strategy. Note that this is only ok because users make the choice to click or not *before* they see different versions of the website (it is "pre-treatment").

Just to make sure you're on track, your measure should have an average value of *about* 9%.

```
ex4_avg_oec = np.mean((df.Enrollments - df.Payments) / df.Clicks)

print(f"The proportion of frustrated users is {ex4_avg_oec:.4f}")

results["ex4_avg_oec"] = np.round(ex4_avg_oec, 4)
```

```
The proportion of frustrated users is 0.0941
```

Answer

Udacity aims to reduce the number of frustrated users—users who enroll in the free trial but do not ultimately pay for the course. Accordingly, the Overall Evaluation Criterion (OEC) is the proportion of frustrated users. This is measured as the difference between enrollments and payments, normalized by the number of clicks.

Normalizing by clicks accounts for day-to-day variation in site traffic and is valid because clicking occurs prior to treatment assignment, making it a pre-treatment variable. This ensures the OEC reflects changes attributable to the experiment rather than fluctuations in user volume.

Exercise 5

Given Udacity's goals, what outcome are they hoping will *not* be impacted by their manipulation? In other words, what do they want to measure to ensure their treatment doesn't have unintended negative consequences that might be really costly to their operation?

Note that while this isn't how Kohavi, Tang, and Xu use the term "guardrail metrics" — they usually use the term to refer to things we measure to ensure the experiment is working the way it should — some people would also use the term "guardrail metrics" for something that could be impacted even if the experiment is working correctly, but which the organization wants to track to ensure they aren't impacted because they are deemed really important.

Again, please normalize by `Clicks`. Store the average value of this guardrail metric as `ex5_avg_guardrail` and **round your answer to 4 decimal places.**

```
ex5_avg_guardrail = np.mean(df.Payments / df.Clicks)

print(f"The proportion of satisfied users is {ex5_avg_guardrail:.4f}")

results["ex5_avg_guardrail"] = np.round(ex5_avg_guardrail, 4)
```

```
The proportion of satisfied users is 0.1158
```

Answer

Udacity hopes that the treatment will not affect the proportion of satisfied users—users who ultimately pay for the course. Accordingly, this guardrail metric is defined as the number of payments normalized by the number of clicks.

While the treatment is intended to discourage less committed users from enrolling in the free trial, it should not deter highly motivated users who are willing to pay. Monitoring this metric ensures that the experiment does not introduce unintended negative effects on revenue-generating users.

Validating The Data

Exercise 6

Whenever you are working with experimental data, the first thing you want to do is verify that users actually were randomly sorted into the two arms of the experiment. In this data, half of users were supposed to be shown the old version of the site and half were supposed to see the new version.

`Pageviews` tells you how many unique users visited the welcome site we are experimenting on. `Pageviews` is what is sometimes called an "invariant" or "guardrail" variable, meaning that it shouldn't vary across treatment arms—after all, people have to visit the site before they get a chance to see the treatment, so there's no way that being assigned to treatment or control should affect the number of pageviews assigned to each group.

"Invariant" variables are also an example of what are known as a "pre-treatment" variable, because pageviews are determined before users are manipulated in any way. That makes it analogous to gender or age in experiments where you have demographic data—a person's age and gender are determined before they experience any manipulations, so the value of any pre-treatment attributes should be the same across the two arms of our experiment. This is what we've previously called "checking for balance," If pre-treatment attributes aren't balanced, then we may worry our attempt to randomly assign people to different groups failed. Kohavi, Tang and Xu call this a "trust-based guardrail metric" because it helps us determine if we should trust our data.

To test the quality of the randomization, calculate the average number of pageviews for the treated group and for the control group. Do they look similar?

```
df.groupby("Treatment").Pageviews.mean()
```

```
Treatment
0    9339.000000
1    9315.135135
Name: Pageviews, dtype: float64
```

Answer

Yes, the average number of Pageviews is very similar across the treatment and control groups, with a difference of less than 0.5%. Because pageviews are a pre-treatment invariant, this similarity suggests that users were randomly assigned as intended. While a formal statistical test could be performed, the small magnitude of the difference provides no immediate reason to suspect a failure of randomization.

Exercise 7

"Similar" is a tricky concept – obviously, we expect *some* differences across groups since users were *randomly* divided across treatment arms. The question is whether the differences between groups are larger than we'd expect to emerge given our random assignment process. To evaluate this, let's use a `ttest` to test the statistical significance of the differences we see.

Note: Remember that scipy functions don't accept `pandas` objects, so you use a scipy function, you have to pass the numpy vectors underlying your data with the `.values` operator (e.g. `df.my_column.values`).

Does the difference in `pageviews` look statistically significant?

Store the resulting p-value in `ex7_ttest_pvalue` **rounded to four decimal places.**

```
feat = "Pageviews"

control = df[df.Treatment == 0][feat].values
tmt = df[df.Treatment == 1][feat].values

t_stats, p_value = stats.ttest_ind(control, tmt, equal_var=False)
print(f"p-value for {feat}: {p_value:.4}")

results["ex7_ttest_pvalue"] = np.round(p_value, 4)
```

```
p-value for Pageviews: 0.8877
```

Answer

Since the p-value is 0.8877 (> 0.05), then we fail to reject $H_0 : \mu_0 = \mu_1$ meaning that there is no statistically significant difference between the amount of `Pageviews` between the Control and Treatment groups.

Exercise 8

`Pageviews` is not the only "pre-treatment" variable in this data we can use to evaluate balance/use as a guardrail metric. What other measure is pre-treatment? Review the description of the experiment if you're not sure.

Answer

`Clicks` is another guardrail metric. As with `Pageviews`, the number of `Clicks` is determined before users are exposed to the treatment, since users must click before seeing either version of the "next" site (pop-up or no pop-up).

Exercise 9

Check if the other pre-treatment variable is also balanced. Store the p-value of your test of difference in `results` under the key `"ex9_ttest_pvalue_clicks"` **rounded to four decimal places.**

```
feat = "Clicks"

control = df[df.Treatment == 0][feat].values
tmt = df[df.Treatment == 1][feat].values

t_stats, p_value = stats.ttest_ind(control, tmt, equal_var=False)
print(f"p-value for {feat}: {p_value:.4}")

results["ex9_ttest_pvalue_clicks"] = np.round(p_value, 4)
```

```
p-value for Clicks: 0.9264
```

Answer

Since the p-value is 0.9264 (> 0.05), then we fail to reject $H_0 : \mu_0 = \mu_1$ meaning that there is no statistically significant difference between the amount of `Clicks` between the Control and Treatment groups, providing evidence that the randomization was successful.

Estimating the Effect of Experiment

Exercise 10

Now that we've validated our randomization, our next task is to estimate our treatment effect. First, though, there's an issue with your data you've been able to largely ignore until now, but

which you should get a grip on before estimating your treatment effect — can you tell what it is and what you should do about it?

Store the number of observations in your data *after* you've addressed this in `ex10_num_obs` (this is mostly meant as a way to sanity check your answer with autograder).

```
print(
    f"We have {df.shape[0]} original observations which correspond to {df.Date.nunique()} uni
)

data = df[~df.Enrollments.isna()].copy()

ex10_num_obs = data.shape[0]

print(
    f"Afterwards, we have {ex10_num_obs} observations without missing values which \ncorrespo
)

results["ex10_num_obs"] = ex10_num_obs
```

```
We have 74 original observations which correspond to 37 unique dates
Afterwards, we have 46 observations without missing values which
correspond to 23 unique dates
```

Answer

The issue is that the dataset contains missing values: while there are 74 observations corresponding to 37 unique dates, 14 of those dates have missing (NA) enrollment/payment data. Because treatment effects are estimated using enrollments/payments, these observations cannot be used and must be excluded. After removing dates with missing values, we are left with 46 observations (23 unique dates) to estimate the treatment effect.

Exercise 11

Now that we've established we have good balance (meaning we think randomization was likely successful), we can evaluate the effects of the experiment. Test whether the OEC and the metric you *don't* want affected have different average values in the control group and treatment group.

Because we've randomized, this is a consistent estimate of the Average Treatment Effect of Udacity's website change.

Calculate the difference in means in your OEC and guardrail metrics using a simple t-test. Store the resulting effect estimates in `ex11_oec_ate` and `ex11_guard_ate` and p-values in `ex11_oec_pvalue` and `ex11_guard_pvalue`. **Please round all answers to 4 decimal places.** Report your ATE in *percentage points*, where 1 denotes 1 percentage point.

```
data["OEC"] = (data.Enrollments - data.Payments) / data.Clicks
data["GUARDRAIL"] = data.Payments / data.Clicks
```

```
feat = "OEC"

control = data[data.Treatment == 0][feat]
tmt = data[data.Treatment == 1][feat]

t_stats, p_value = stats.ttest_ind(control, tmt)

ex11_oec_pvalue = p_value
print(f"p-value for {feat}: {p_value:.4}")

print(
    f"\nThe average % of unsatistied users, per Click, \nfor the Treatment group is {100 * np
)
print(
    f"The average % of unsatistied users, per Click, for the Control group is {100 * np.mean
)

ex11_oec_ate = np.mean(tmt) - np.mean(control)
ex11_oec_ate = np.round(100 * ex11_oec_ate, 4)
print(
    f"\nDifference in means for the {feat} (% of unsatisfied\nusers, per Click, between Treat
)

results["ex11_oec_ate"] = np.round(ex11_oec_ate, 4)
results["ex11_oec_pvalue"] = np.round(ex11_oec_pvalue, 4)
```

```
p-value for OEC: 0.1319

The average % of unsatistied users, per Click,
for the Treatment group is 8.6194
The average % of unsatistied users, per Click, for the Control group is 10.2082

Difference in means for the OEC (% of unsatisfied
users, per Click, between Treatment and Control): -1.5888 percentage points
```

Since the p-value is 0.132 (> 0.05), then we fail to reject $H_0 : \mu_0 = \mu_1$ meaning that there is no statistically significant difference between the average `OEC` between the Control and Treatment groups.

```
feat = "GUARDRAIL"

control = data[data.Treatment == 0][feat]
tmt = data[data.Treatment == 1][feat]

t_stats, p_value = stats.ttest_ind(control, tmt, equal_var=False)

ex11_guard_pvalue = np.round(p_value, 4)
print(f"p-value for {feat}: {p_value:.4}")

print(
    f"\nThe average % of satistied users, per Click, \nfor the Treatment group is {100 * np.m
)
print(
    f"The average % of satistied users, per Click, for the Control group is {100 * np.mean(co
)

ex11_guard_ate = np.mean(tmt) - np.mean(control)
ex11_guard_ate = np.round(100 * ex11_guard_ate, 4)
print(
    f"\nDifference in means for the {feat} (% of satisfied\nusers, per Click, between Treatme
)

results["ex11_guard_ate"] = np.round(ex11_guard_ate, 4)
results["ex11_guard_pvalue"] = np.round(ex11_guard_pvalue, 4)
```

```
p-value for GUARDRAIL: 0.5928

The average % of satistied users, per Click,
for the Treatment group is 11.3373
The average % of satistied users, per Click, for the Control group is 11.8269

Difference in means for the GUARDRAIL (% of satisfied
users, per Click, between Treatment and Control): -0.4897 percentage points
```

Since the p-value is 0.5928 (> 0.05), then we fail to reject $H_0 : \mu_0 = \mu_1$ meaning that there is no statistically significant difference between the average `Guardrail` metrics between the Control and Treatment groups.

Exercise 12

Do you feel that Udacity achieved their goal? Did their intervention cause them any problems? If they asked you "What would happen if we rolled this out to everyone?" what would you say?

As you answer this question, a small additional question: up until this point you've (presumably) been reporting the default p-values from the tools you are using. These, as you may recall from stats 101, are two-tailed p-values. Do those seem appropriate for your OEC?

Answer

There is no statistically significant difference in the OEC between the control and treatment groups. This means we don't have evidence that the treatment reduced the proportion of frustrated users. The guardrail metric for satisfied users also shows no significant change, so the intervention didn't cause any measurable problems.

The default p-values from the t-tests are two-tailed, which test for any difference, either an increase or a decrease in OEC ie. $H_0 : \mu_0 = \mu_1$ vs. $H_1 : \mu_0 \neq \mu_1$. Udacity's goal is directional since they want OEC to decrease, so a one-tailed test could also be considered ie. $H_0 : \mu_0 = \mu_1$ vs. $H_1 : \mu_0 > \mu_1$ where μ_0 is the mean OEC for the control group. If we calculate a one-tailed p-value, it would be approximately $0.132/2 = 0.066$, which is still above the 0.05 threshold. This confirms that even when focusing only on the expected direction, the treatment effect is not statistically significant.

If Udacity rolled this change out to everyone, we would likely see no effect on the OEC. From a business perspective, this means the website change does not appear to have a meaningful impact. If they want stronger evidence, the experiment could be rerun with more users or more days to increase statistical power.

Side note

Note: If Udacity were willing to use a more permissive significance threshold (e.g., $= 0.10$) and focus on a one-tailed test aligned with the expected direction of the effect, the result would be marginally significant (p 0.066). Whether acting on this evidence is appropriate depends on the relative costs of false positives versus false negatives. If the intervention is low-risk and inexpensive to deploy, Udacity might reasonably accept weaker statistical evidence; if the costs of a mistaken rollout are high, the current evidence would likely be insufficient.

Exercise 13

One of the magic things about experiments is that all you have to do is compare averages to get an average treatment effect. However, you *can* do other things to try and increase the statistical power of your experiments, like add controls in a linear regression model.

As you likely know, a bivariate regression is exactly equivalent to a t-test, so let's start by re-estimating the effect of treatment on your OEC using a linear regression. Can you replicate the results from your t-test? They shouldn't just be close—they should be numerically equivalent (i.e. exactly the same to the limits of floating point number precision).

```
model = smf.ols("OEC ~ Treatment", data).fit()
model.summary()
```

Dep. Variable:	OEC	**R-squared:**	0.051
Model:	OLS	**Adj. R-squared:**	0.029
Method:	Least Squares	**F-statistic:**	2.356
Date:	Sat, 31 Jan 2026	**Prob (F-statistic):**	0.132
Time:	22:26:59	**Log-Likelihood:**	89.832
No. Observations:	46	**AIC:**	-175.7
Df Residuals:	44	**BIC:**	-172.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.1021	0.007	13.948	0.000	0.087	0.117
Treatment	-0.0159	0.010	-1.535	0.132	-0.037	0.005

Omnibus:	14.160	**Durbin-Watson:**	1.908
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	15.205
Skew:	1.227	**Prob(JB):**	0.000499
Kurtosis:	4.383	**Cond. No.**	2.62

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Answer

The coefficient for the treatment is -0.0159, which is the same result as we got in the last exercise (12) with the t-test of the difference in means for the OEC. In a randomized experiment, estimating a regression of OEC on a treatment indicator yields the same average treatment effect as comparing group means.

The coefficient implies that the treatment reduces the OEC by approximately 1.59 percentage points on average, corresponding to a reduction in the share of unsatisfied users. However, consistent with the t-test results, this effect is not statistically significant.

Exercise 14

Now add indicator variables for the date of each observation. Do the standard errors on your `treatment` variable change? If so, in what direction?

Store your new standard error in `ex14_se_treatment`. Round your answer to 4 decimal places.

```
model_14 = smf.ols("OEC ~ Treatment + C(Date)", data).fit()
model_14.summary()
```

Dep. Variable:	OEC	**R-squared:**	0.806
Model:	OLS	**Adj. R-squared:**	0.602
Method:	Least Squares	**F-statistic:**	3.962
Date:	Sat, 31 Jan 2026	**Prob (F-statistic):**	0.000978
Time:	22:26:59	**Log-Likelihood:**	126.29
No. Observations:	46	**AIC:**	-204.6
Df Residuals:	22	**BIC:**	-160.7
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	0.1079	0.016	6.651	0.000	0.074	0.142
C(Date)[T.Fri, Oct 24]	0.0445	0.022	1.983	0.060	-0.002	0.091
C(Date)[T.Fri, Oct 31]	-0.0074	0.022	-0.331	0.744	-0.054	0.039
C(Date)[T.Mon, Oct 13]	-0.0231	0.022	-1.026	0.316	-0.070	0.024
C(Date)[T.Mon, Oct 20]	-0.0285	0.022	-1.270	0.217	-0.075	0.018
C(Date)[T.Mon, Oct 27]	0.0328	0.022	1.458	0.159	-0.014	0.079
C(Date)[T.Sat, Nov 1]	-0.0235	0.022	-1.047	0.306	-0.070	0.023
C(Date)[T.Sat, Oct 11]	-0.0017	0.022	-0.074	0.941	-0.048	0.045
C(Date)[T.Sat, Oct 18]	-0.0438	0.022	-1.950	0.064	-0.090	0.003
C(Date)[T.Sat, Oct 25]	-0.0309	0.022	-1.375	0.183	-0.077	0.016
C(Date)[T.Sun, Nov 2]	0.0549	0.022	2.441	0.023	0.008	0.101
C(Date)[T.Sun, Oct 12]	-0.0347	0.022	-1.542	0.137	-0.081	0.012
C(Date)[T.Sun, Oct 19]	-0.0178	0.022	-0.791	0.437	-0.064	0.029
C(Date)[T.Sun, Oct 26]	-0.0222	0.022	-0.989	0.333	-0.069	0.024
C(Date)[T.Thu, Oct 16]	-0.0228	0.022	-1.016	0.321	-0.069	0.024
C(Date)[T.Thu, Oct 23]	-0.0046	0.022	-0.203	0.841	-0.051	0.042
C(Date)[T.Thu, Oct 30]	0.0588	0.022	2.618	0.016	0.012	0.105
C(Date)[T.Tue, Oct 14]	-0.0417	0.022	-1.855	0.077	-0.088	0.005
C(Date)[T.Tue, Oct 21]	-0.0059	0.022	-0.260	0.797	-0.052	0.041
C(Date)[T.Tue, Oct 28]	-0.0287	0.022	-1.276	0.215	-0.075	0.018
C(Date)[T.Wed, Oct 15]	-0.0132	0.022	-0.588	0.562	-0.060	0.033
C(Date)[T.Wed, Oct 22]	-0.0220	0.022	-0.980	0.338	-0.069	0.025
C(Date)[T.Wed, Oct 29]	0.0466	0.022	2.076	0.050	4.77e-05	0.093
Treatment	-0.0159	0.007	-2.398	0.025	-0.030	-0.002

Omnibus:	3.871	**Durbin-Watson:**	1.863
Prob(Omnibus):	0.144	**Jarque-Bera (JB):**	3.826
Skew:	-0.000	**Prob(JB):**	0.148
Kurtosis:	4.413	**Cond. No.**	27.3

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
se_model = model.bse["Treatment"]
se_model_14 = model_14.bse["Treatment"]

print(
    f"The standard errors for the model that only considers the effects \nof OEC on the Treat
)
print(
    f"\nThe standard errors for the model that considers the effects \nof OEC + dates on the
)
```

```
print(
    f"\nThe difference between both standard errors is {100 * ((se_model - se_model_14) / se
)
```

```
The standard errors for the model that only considers the effects
of OEC on the Treatment is 0.0104

The standard errors for the model that considers the effects
of OEC + dates on the Treatment is 0.0066

The difference between both standard errors is 35.99%
```

```
ex14_se_treatment = round(model_14.bse["Treatment"], 4)
results["ex14_se_treatment"] = ex14_se_treatment
```

Answer

By adding date indicators, the standard errors decrease by about 36% compared to the model that only includes the treatment variable (from 0.0104 to 0.0066). Therefore, the standard error on the treatment variable becomes smaller when controlling for date fixed effects. The p-value is now 0.025 (<0.05), indicating that the treatment effect is statistically significant.

The treatment coefficient remains the same as before (-0.0159), but including date controls increases the precision of the estimate.

You should have found that your standard errors decreased by about 30%—this is why, although just comparing means *works*, if you have additional variables adding them to your analysis can be helpful (all the usual rules for model specification apply — for example, you still want to be careful about overfitting, which one could argue is maybe part of what's happening here).

In many other cases, the effect of adding controls is likely to be larger — the date indicators we added to our data are perfectly balanced between treatment and control, so we aren't adding a lot of data to the model by adding them as variables. They're accounting for some day-to-day variation (presumably in the types of people coming to the site), but they aren't controlling for any residual baseline differences the way a control like "gender" or "age" might (since those kind of individual-level attributes will never be perfectly balanced across treatment and control).

Exercise 15

Does this result have any impact on the recommendations you would offer Udacity?

Answer

Yes, this result affects the recommendation, but it does not fully overturn the original conclusion. Adding date fixed effects increases the precision of the treatment estimate and renders the effect statistically significant, which strengthens confidence that the treatment may reduce the OEC.

However, because the original difference-in-means analysis did not yield a statistically significant result and the inclusion of date controls was not part of the original experimental design, this evidence alone may not be sufficient to justify a full rollout. A reasonable recommendation would be to rerun the experiment with greater power or pre-specify the use of date fixed effects before making a final deployment decision.

```
assert set(results.keys()) == {
    "ex4_avg_oec",
    "ex5_avg_guardrail",
    "ex7_ttest_pvalue",
    "ex9_ttest_pvalue_clicks",
    "ex10_num_obs",
    "ex11_guard_ate",
    "ex11_guard_pvalue",
    "ex11_oec_ate",
    "ex11_oec_pvalue",
    "ex14_se_treatment",
}
```